Data analysis techniques: a tool for cumulative exposure assessment.

Benoît Lalloué, Jean-Marie Monnez, Cindy Padilla, Wahida Kihal, Denis Zmirou-Navier, Séverine Deguen

Title: Data Analysis Techniques: a Tool for Cumulative Exposure Assessment

Running title: Data Analysis for Cumulative Exposure Index

Benoît Lalloué, PhD, 1,2,3,4

Jean-Marie Monnez, PhD, 3,4

Cindy Padilla, PhD, 1,2

Wahida Kihal, PhD, 1,2

Denis Zmirou-Navier, MD, PhD, 1,2,5

Séverine Deguen, PhD, 1,2

1 EHESP Rennes, Sorbonne Paris Cité, France

2 Inserm, UMR1085-IRSET (Institut de recherche sur la santé l'environnement et le travail), France

3 Lorraine University, Institut Elie Cartan de Lorraine, CNRS UMR 7502, France

4 Lorraine University, INRIA, CNRS UMR7502, BIGS (INRIA Nancy - Grand Est / IECL), France

5Lorraine University Medical School, France

Corresponding author:

Séverine Deguen

EHESP, Avenue du Professeur Léon Bernard,

CS 74312 - 35043 Rennes cedex

severine.deguen@ehesp.fr

Tel : +33299022805

Fax : +33299022675

Financial support: This work and the Equit'Area project are supported by the French National Research Agency (ANR, contract-2010-PRSP-002-01) and the EHESP School of

Public Health. This research was also jointly supported by the Direction Générale de la Santé (DGS), the Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS), the Régime Social des Indépendants (RSI), the Caisse Nationale de Solidarité pour l'Autonomie (CNSA), the Mission Recherche de la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (MiRe-DREES) and l'Institut national de prévention et de promotion de la santé (Inpes), under the research call launched by the French Institute of Public Health Research (IReSP) in 2010.

ABSTRACT

Background: Everyone is subject to environmental exposures from various sources, with negative health impacts (air, water and soil contamination, noise ...) or with positive effects (e.g. green space). Studies considering such complex environmental settings in a global manner are rare. We propose to use statistical factor and cluster analyses to create a composite exposure index with a data-driven approach, in view to assess the environmental burden experienced by populations. We illustrate this approach in a large French metropolitan area.

Methods: The study was carried out in the Great Lyon area (France, 1.2M inhabitants) at the census block group (BG) scale. We used as environmental indicators ambient air NO_2 annual concentrations, noise levels, proximity to green spaces, to industrial plants, to polluted sites and to road traffic. They were synthesized using Multiple Factor Analysis (MFA), a data driven technique without *a priori* modeling, followed by a hierarchical clustering to create BG classes.

Results: The first components of the MFA explained respectively 30, 14, 11 and 9% of the total variance. Clustering in 5 classes group: 1) A particular type of large BGs without population; 2) BGs of green residential areas, with less negative exposures than average;3) BGs of residential areas near midtown; 4) BGs close to industries; 5) midtown urban BGs, with higher negative exposures than average and less green spaces. Other numbers of classes were tested in order to assess a variety of clustering.

Conclusions: We present an approach using statistical factor and cluster analyses techniques which seem overlooked to assess cumulative exposure in complex environmental settings. Although it cannot be applied directly for risk or health effect assessment, the resulting index

can help to identify hot spots of cumulative exposure, to prioritize urban policies or to compare the environmental burden across study areas in an epidemiological framework.

KEYWORDS

Cumulative exposure; Environmental index; Multiple Factor Analysis; Hierarchical Clustering

INTRODUCTION

At each moment, everyone is subject to a mixture of environmental exposures in one's place of residence. Of these, air pollution (1,2), water and soil contamination (3), noise (4,5) and proximity to garbage dumps or hazardous industrial facilities (3) are separately recognized to have adverse health effects such as respiratory and heart diseases, cancer, adverse pregnancy outcomes or mental health impairment (1–5). Conversely, other environmental exposures, such as proximity to green spaces, have demonstrated positive health effects (6,7).

The majority of epidemiological studies have considered only one single environmental exposure when assessing its possible health impact. Recently, there has been a call for more 'realistic' approaches that would take into account complex exposure situations, and strive to measure the 'environmental burden' experienced by populations (8–15). New tools to assess cumulative exposures would be useful in a wide range of applications, from public policy and urban planning to environmental epidemiology or environmental justice studies.

One reason making it difficult to consider cumulative exposures is that several methodological issues arise in dealing with them. Exposure variables are often correlated since they can share the same sources of pollution (such as road traffic for instance), which can imply multicollinearity issues. Also, the different types of variables used to characterize the quality of the environmental media, having both qualitative (presence/absence indicators for instance) and quantitative (pollution levels) indicators, as well as the various units of measurement ($\mu g/m^3$, decibel, g/m^2, m^2, percentage of population exposed above a certain level, etc.), constitute additional difficulties in constructing an index of cumulative environmental exposure. Finally, determining the weighting of the different exposure

variables included in a composite index is also complex - particularly when each exposure is associated with different health outcomes.

To our knowledge, the few studies that took into account cumulative exposures often included pollutants of the same family (mostly air pollution) and/or used simple methods such as weighted sum or product, or arbitrarily defined scores (9,16–19).

One solution to these methodological limitations is to use data analysis techniques which can include various types or scales of variables from large data sets, and which are underemployed in this particular context. Moreover, these techniques are data driven, in the sense that they allow the data to organize itself without *a priori* knowledge or model, and their analysis may highlight the underlying relations and correlations between variables.

In this context, the aim of the present work is to construct an index of complex cumulative environmental exposure that allows to assess the 'environmental burden' experienced by the population in a French metropolitan area. To do so, we chose to explore the usefulness of a data driven approach with no *a priori* model and to create this index independently of any outcome variable, in view to highlight the underlying structure of our environmental data. This work is a part of the Equit'Area research project (www.equitarea.org), whose main objective is to study the combined impact of multiple environmental exposures and of the contextual socio-economic status on the risk of infant and neonatal mortality.

MATERIAL AND METHODS

Study setting and small area level

The study was carried out in the Lyon Metropolitan Area (LMA) known as the 'Grand Lyon', which is one of the biggest such areas in France and is located in the Rhône-Alpes region (central and eastern France). The LMA is subdivided into 56 municipalities and 499 census

Block Groups (BG), for a total population of approximately 1.2 million inhabitants in an area of 527 km².

The statistical units are the sub-municipal French BGs (called IRIS), defined by the National Institute of Statistics and Economic Studies (INSEE(20)). In the whole of France, these units have an average of 2,000 inhabitants and are constructed so as to be as homogeneous as possible in terms of socio-demographic characteristics and land use. They also take account of physical obstacles that may break up urban landscapes, such as arterial roads, green spaces, bodies of water. In the LMA, BGs ranged from 0 to 7,232 inhabitants, with an average 2,465 inhabitants in 2007.Their area varied from 0.01 to 11.15km² (1.00 km² on average).

Environmental variables

The selection of environmental exposures was based on their known or suspected links (positive or negative) with the study health outcome, from the literature, as well as on the availability of databases needed to construct the environmental indicators at the BG level at the time of the study. Others exposure factors that may be relevant in the LMA, such as particulate matter or ozone, could not be obtained at the BG level at the time of the study. In other settings, such factors might be added in the cumulative exposure index construction whenever available.

For each of the following exposure types, several indicators were available and are described below. Descriptive statistics of these indicators at the BG level are given only for those chosen by the procedure (see below: "Methodological Procedure, Selection of variables"). The Pearson correlations between these indicators are presented in Table 1. Indicators of the same exposure domain are unsurprisingly strongly correlated, as well as indicators of different exposure domains sharing a common source (such as traffic exhausts and noise).

Nitrogen dioxide

Annual averages of nitrogen dioxide (NO_2) concentrations (expressed in $\mu g/m^3$) were modelled at a 10 x 10m resolution and then aggregated at the census Block Group (BG) level throughout the entire study period (2002-2009) by the local air monitoring association (Air Rhône-Alpes) using the SIRANE modelling system (21,22). More details about estimated NO_2 data are available elsewhere(23).

Across the LMA, the mean annual average of NO_2 concentration was respectively 40.31 and 42.44 $\mu g/m^3$ in 2003 and 2006 (table 2), which were in both cases above the exposure threshold set in the European Union for yearly values ($40\mu g/m^3$)(24).

Noise

Exposure to residential noise combined noise nuisances related to road and aircraft traffic, industries and railway. Noise nuisances were measured in 2007 and the acoustic modelling estimations of noise levels, with a spatial resolution of 10×10 meters at 4 metres above ground level, across the LMA were performed by the Grand Lyon Urban Community (the political institution regrouping the municipalities of the LMA) in line with the European Environmental Noise Directive (25).

The metric used to characterize noise in each census block was the European standard L_{den} indicator (day-evening-night Level, measured in decibels, dB), an assessment of daily exposure over a 24-hour period taking into account residents' increased sensitivity to noise during the evening (6 pm to 10 pm) and night (10 pm to 6 am) (5). Combining the L_{den} with the estimated population living in each residential building, the Scientific and Technical Centre for Building (CSTB) calculated several indicators to compute a (weighted or not) population average noise exposure at the BG level. More details about how the indicators were constructed can be found elsewhere (26).

Noise levels had a median BG arithmetic mean of 63.85 dB(A) without taking the buildings' population into account, and 65.58 dB(A) when it is (table 2). In France, the regulatory L_{DEN}

threshold fixed in application of the European Environmental Noise Directive (25) is 68 dB(A). This threshold was exceeded in 43 BGs (8.7%) for the arithmetic mean, and 126 BGs (25.6%) when weighing on the population size.

Proximity to industrial plants and polluted sites

Proximity to industrial plants was determined using data from the European Pollutant Emission Register (EPER register)(27) which routinely collects data from industries emitting pollutants above a fixed emission threshold for about 50 different pollutants. Proximity to polluted sites were determined using data from the BASIAS database (28) -a French register of all former industrial plants and service activities where polluted installations or polluted soils might remain.

Twenty six industrial plants registered in the EPER database and located in the metropolitan area were included in this study, as well as 474 polluted sites in the area registered in BASIAS database. The geographical coordinates for each of the selected plants were checked and corrected (where necessary) using Google Maps.

Several indicators were created at the BG level using the ArcGIS software (29) to assess proximity to industrial plants or polluted sites.

For industrial plants: presence/absence of a pollutant industry within the BG; number of industries emitting pollutants within the BG; presence/absence of a buffer with a 500m or 1km radius around an industrial plant intersecting with the BG; number of buffers with a 500m or 1km radius around an industrial plant intersecting with the BG.

For polluted sites: number of sites within the BG; number of buffers with a 500m, 1km or 1.5km radius around a polluted site intersecting with the BG.

Due to the low number of industrial plants in the area, indicators related to this exposure were all considered qualitative, whereas indicators related to the polluted sites were considered quantitative since there are enough polluted sites in each BG to do so.

More than 70% of BG in the metropolitan area are more than 1km away from industrial plants. However, 21 BG (4.2%) were within 500m of at least 2 industrial plants (table 2). There was also an average of 7 buffers of 500m radius around polluted sites intersecting with a BG (table 2).

Traffic exposure

Road traffic was assessed using the Grand Lyon Urban Community and Air Rhône-Alpes traffic model. 'High-traffic' road sections were defined as those having more than 5,000 vehicles per day. Using the estimated traffic for each road section, combined with information about buildings and population, the proportion of the population within a given distance of a 'high-traffic' section was computed for each BG.

We constructed and tested these indicators with several strip widths (from 100m to 300m, in 50m steps).

An average of 77% of the BG population was within 200m of a 'high-traffic' road. However, since the entire range 0-100% was covered, there were extremely broad variations between BGs (table 2).

Green Spaces

Spatial land cover data sets from the French National Geographic Institute (IGN) were sought and processed using ArcGIS software for the production of a green spaces index. Our definition of green space (30) included natural areas (e.g., parks, forest…). This green spaces index was defined as the proportion of the geographic area occupied by green spaces within the total area of a census block. The total green space area for each BG is also included. Green spaces presented wide ranges of variations in the LMA, with BGs having 0 to 57% of their area occupied by green spaces (table 2).

Spatial patterns of the environmental indicators

From a spatial point of view, both NO_2 levels (Figure 1, panel a.) and road traffic proximity (Figure 1, panel b.) were greater in the city of Lyon itself and around north-south highway crossing the city (the "*Autoroute du soleil"*, connecting Paris to the French Mediterranean coast), whereas noise levels (Figure 1, panel c.) were more mixed across the metropolitan area, with no clear spatial pattern.

Both polluting industries (Figure 1, panel d.) and polluted sites (Figure 1, panel e.) were mainly located in the east, south-east and south of the city of Lyon and the greenest BGs (in absolute value) in the metropolitan area were on the outskirts, in contrast with the city centre BGs. However, looking at green space as a proportion of the BG area (Figure 1, panel f.), a clear heterogeneity appears between the north-western and the south-eastern BGs the latter having a lower percentage of green spaces.

Methodological Procedure

Selection of variables

In total, we had 31 environmental indicators divided into the six environmental groups previously described: air pollution, noise, industrial proximity, traffic, polluted sites, and green spaces. With the notable exception of the industrial proximity group which is qualitative, each of these is composed of quantitative variables. Since there are correlations within the groups we wanted to select a subset of indicators for each group in order to limit the number of indicators while keeping enough information. To do so, we used principal component analysis (PCA) for each group of quantitative variables or multiple correspondence analysis (MCA)(31) for the industrial proximity group (results not shown), keeping only the 1 or 2 indicators most correlated with the first and second components (similar, in a way, to the first step of the procedure of creation of a composite socio-economic index developed by our team and published elsewhere(32)). The tested but eventually dropped

indicators were: (1) 2002, 2004, 2005 and 2007-2008 NO_2 annual concentration average; (2) Noise levels energetic mean and noise levels median taking into account the buildings population; (3) Presence of at least one pollutant industry and number of 1km radius buffers around an industrial plant intersecting with the BGs; (4) Proportion of the BG population within 100m, 150m or 300m of a high traffic road; and (5) Number of 1500m radius buffers around a polluted site intersecting with the BGs. We then obtained a selection of 17 environmental indicators (see tables2 and 3).

Data analysis techniques

In this study, our purpose was to create cumulative exposure index based on the underlying structure of the data with a data driven approach and without any *a priori* model. Since we had several groups of both quantitative and qualitative environmental indicators (air pollution, noise, industrial proximity, traffic, polluted sites, green spaces) and because we wanted to give an equal weight to each, regardless of the number of indicators in it, we used multiple factor analysis (MFA)(33) which is a well-suited technique for this situation:

Consider a data set composed of observations among the same units of several groups of variables. This dataset can be divided in subsets representing the groups of variables.

The first step of the MFA is, for each subset, to perform a principal component analysis (PCA) if the group is composed of quantitative variable or a multiple correspondence analysis (MCA) if the group is composed of qualitative (dummy) variables. This first step allows to determine a particular metric (i.e. the way to compute distance between units by giving a particular weight to each variable), based on the use of the highest eigenvalue of the PCA or the MCA of each group, which will allow to give in some sense the same weight to each group even if they are of very different sizes.

The second step of the MFA is to perform a PCA on the whole data set, using the previously obtained metric. This allows the comparison of groups of different types of variables.

The interpretation of the MFA is then similar to a PCA: new variables, uncorrelated and of maximal variance, are created as linear combinations of the original variables (without any factors rotation) and can be interpreted using correlations with original variables or contributions of variables and groups together with a set of graphical outputs which facilitate interpretation.

Following the MFA, we applied Hierarchical Clustering (HC) to create meaningful categories of BGs. HC is a unsupervised method of clustering which creates a hierarchy of classes (i.e. clusters), frequently used after a PCA or others data analysis techniques such as MFA.

Given a set of variables (here, these variables are those created by the MFA) the HC algorithm creates a hierarchy of categories step by step, at each step merging the two categories which are closest, according to a given distance between categories. When it is a particular distance (the Ward distance), this algorithm aims to obtain categories that are homogeneous within and heterogeneous between one another with respect to an inertia-based criterion.

The most appropriate partition is then selected from the hierarchy of categories, using both mathematical criterion (such as the dendrogram) and knowledge (in order to keep a reasonable number of meaningful categories). Each category can then be interpreted according to the average values of the quantitative variables for the BG and/or to the proportions of the different modalities of the qualitative variables in this category, as well as according to the BGs which compose it. More information about PCA, MFA and HC is available in the supplementary Text S1.

This method allows to create a categorical index of cumulative exposure with a data driven approach and simple data analysis techniques (see Figure 2). Note that due to the data driven nature of the techniques used here, this process should be repeated entirely in each new setting. A main advantage is that even if one has not the same variables characterising the environmental exposures, the same procedure can still be applied. All statistical and data analysis computation has been conducted using the R software (34) and the package FactoMineR (35).

RESULTS

Multiple Factor Analysis

The MFA was applied on the 17 selected variables covering the 6 exposure groups described above. The four first components explain 30%, 14%, 11% and 9% respectively of the total variance (supplementary Table S1). These components can be interpreted using both the groups and the variables contributions (see Table 4 and supplementary Table S2) to the components or their graphical representations (supplementary Figures S1 and S2):

(1) The first component can be expressed as an air pollution and traffic proximity component, which were two groups of strongly correlated families of indicators (Table 1). BGs with high values in this component are mainly located in the centre of the LMA (supplementary Figure S3). (2) The second component is mainly explained by industrial proximity and is clearly based on the opposition between the presence and the absence of industries or buffers in BGs. BGs with high values in this component are mostly in the south-eastern part of the LMA but no clear pattern appears. (3) The third component relates to noise and green spaces, which is more surprising according to the small two-by-two correlations between the indicators. It is positively correlated both with green spaces indicators and with noise variables (particularly those which do not take into account exposed population sizes); most BGs with high values in this component are in the north-western part of the LMA. (4) Finally, the fourth component

relates to noise and polluted sites, which was also not expected in view of the correlations. Noise indicators which take into account the buildings' population have a strong negative correlation with this component. No clear geographical pattern appears for this component.

Hierarchical Clustering

Following the MFA, we performed an HC on the first five components of the MFA and, according to both the dendrogram and the number of categories expected, we chose a 5-category partition. Using the characteristics of each category according to the variables (Tables 5 and 6), different exposure profiles can be identified in the LMA (Figure 1).

The major characteristic of Category 1 (in red in Figure 1, panel g.) is that all its BGs have a value of 0 for noise variables, taking into account buildings' population. An in-depth examination of the BGs comprising this category revealed that these are wide, sparsely populated areas made of parks, forests, industrial estates or business districts- and present no spatial pattern.

Category 2 (dark green) is the greenest category, with green spaces covering an area ten times the size of the LMA average. BGs in this category also present lower noise levels and NO_2 concentrations, as well as traffic proximity and polluted sites indicators that are well below the study area average. BGs included in this category appear to be residential areas with individual houses and gardens, or rural areas with detached houses and fields. This category is located mainly on the outskirts, especially in the north-west area, of the LMA.

Category 3 (light green) has, on average, lower values of NO_2 and noise levels, of number of polluted soils and of traffic proximity. Although their absolute green space area is smaller than on average in the study area, BGs in this category are proportionally greener than average for their size. BGs in this category also appear to be residential areas with individual houses and gardens, though closer to city centres and smaller than those in Category 2. These

are mainly located on the south-east and midway between the metropolitan area's centre and outskirts.

Category 4 (blue) is characterized by its proximity to pollutants-emitting industries. All BGs in this category are within 500m of an emitting industry and this category contains almost all BGs having a non-0 modality for any industrial proximity variable. This category of BGs near industries is mainly located on the mid-distance outskirts of the LMA.

Finally, Category 5 (orange) BGs have higher noise and NO_2 levels, higher traffic and polluted soils indicators and below-average areas of green space. This category essentially comprises the Lyon municipality itself and neighbouring towns.

Although the construction of the index aims to create a qualitative and nominal index, we can see in its application that a hierarchy emerges between the categories. Category 2 seems to be greener and less exposed to air, noise and traffic pollution, whereas Category 5 carries a significant environmental burden comprising air, noise and traffic pollution as well as close proximity to polluted sites, and a very small relative surface of green spaces.

Examples: links with SES and with infant mortality

In order to illustrate the utility of our index, we present two examples exploring social and spatial inequalities in exposure (36) (i.e. different socio-economic groups bearing an unequal environmental burden).Then, we compare the infant mortality rate and the Socio-Economic Status (SES) of LMA BGs according to their exposure category. To do this, we used a SES index defined elsewhere (32) by our team, which synthesizes around 20 socio-economic variables into a single indicator, using successive principal component analyses (the higher the SES index, the more deprived the BG). The infant mortality rate was calculated using 2002-2009 infant mortality cases gathered in the death registries of the LMA municipalities and the number of living births from the French National Institute of Statistical Studies (INSEE).

Comparing average SES index values by exposure category, we can see (Figure 3) that environmental inequalities are present in the LMA. For instance, Category 2 (green BGs with low air and noise pollution) has, on average, fewer deprived BGs (average SES index: -1.3) whereas Category 4 (BGs closed to pollutant industries) has more deprived BGs (average SES index: 0.7). More generally, BGs in categories with the lowest exposures to NO_2, noise or traffic (categories 1 to 3) are less deprived than those more exposed to NO_2 or pollutant industries (categories 4 and 5). A Kruskal-Wallis test performed on this data confirms a statistically significant difference in SES index distributions between cumulative exposure categories ($p<10^{-6}$).

Comparing infant mortality rates distributions across cumulative exposure categories, we observed a significant link ($p<10^{-15}$ using a Kruskal-Wallis test). Although the very small average rate in category 1 (2.10‰) cannot be interpreted because both of the small number of BGs in this category and of their small population size, one observes that category 2 (the least exposed) presents smaller infant mortality rates than categories 3, 4 and 5. Associations between infant mortality rates and specific exposures split into quintiles (Table 7) showed not significant (NO_2, p=0.18; noise, p=0.65) or were less significant (green spaces, $p<10^{-2}$).

DISCUSSION

In this article, we present a cumulative exposure index aimed to assess the 'environmental burden' among a population. We used statistical data analysis techniques - Multiple Factor Analysis and Hierarchical Clustering in particular - to create this index in a data-driven manner. We give an example of its application in the Lyon Metropolitan area and synthesize exposures such as air pollution, noise, industrial, polluted sites and traffic proximity and green spaces, allowing to obtain a classification of the census Block Groups in 5 easily-interpretable categories representing the diversity of exposure profiles across the metropolitan area.

Studies considering exposures of different types and natures (defined as a ‘coincidental mixture’ by Sexton and Hattis(37)) in order to measure an ‘environmental burden’ and particularly using this type of data analysis techniques, appear to be thin on the ground. It is therefore difficult to compare ours to the literature.

Among the few articles mentioning statistical data analysis techniques for this purpose, Menzie et al (38) cite Principal Component Analysis (PCA) and Cluster Analysis as some of the many statistical methods capable of addressing the health impact of multiple exposures within a general context. Similarly, Billionnet et al (39) also listed PCA, Supervised PCA and Hierarchical Clustering as interesting methods for assessment of the health impact of multiple air pollutants (indoor or outdoor).

When studying the links between polychlorinated biphenyls (PCBs) and hypertension, Christensen and White (18) used Cluster Analysis and Discriminant Analysis in order to identify clusters of similar PCBs and PCA to construct new variables without multi-collinearity for inclusion in regression models. However, this study focuses only on exposures from the same (broad) chemical family measured at individual level and does not consider a neighbourhood environmental burden.

In 2013, Benmarhnia et al (40) proposed a Spatial Environment Index including the quantitative environmental data that is routinely collected at French departmental level (air pollution, water pollution, industrial risks, noise and housing conditions) using a series of PCAs. Moreover, they used certain data, not available at the BG level, which is incapable of assessing environmental particularities at a fine geographic level such as BGs. Therefore, once the departments have been placed in order of priority, there is still a need for more refined detail, which can be obtained using an approach like the one we propose here. Indeed,

the BG level appears to be better, both when studying environmental inequities and for reducing the ecological bias(41).

In the United Kingdom and New Zealand, Richardson et al (17) and Shortt et al (9,42) have developed the MEDIxand the NZ-MEDIx, which are multiple environmental deprivation indices based on scores at small-area levels for environmental exposures such as air pollution, climate (average temperature), proximity to polluting industries, UV radiation and green spaces. The same team also developed the MEDClass and the NZ-MEDClass qualitative indices which "characterize areas according to their shared physical environmental features"(42) with the same exposures as MEDIx and NZ-MEDIx. To do so, they used statistical data analysis techniques both for reducing correlations between UV exposition and temperature, followed by a two-step clustering in order to categorize the different environmental profiles. Associations between both MEDIx and MEDClass (and their New-Zealand equivalents) and health events, such as all-cause mortality, cardiovascular or respiratory diseases, have been shown. For instance, Pearce et al (9) show a statistically significant difference in the all-cause mortality SMR between different environmental categories, the lowest SMR being in the least environmentally deprived areas.

Our approach shares many similarities with the MEDClass construction. However, we perform the MFA on the whole set of data rather restricting it to certain categories, giving the same weight to each group of exposures, and applying the HC directly to the MFA components. Using MFA also allows to include qualitative variables in the analysis, and to extend the set of variables which can be taken into account in assessing the environmental exposure profiles in a given area.

The main thrust of our study lies in usage of data analysis techniques with a data-driven approach, which have been underemployed in environmental health in general and in the cumulative exposure assessment field in particular. After inclusion of the available variables

and definition of the groups in the MFA, this approach allows to obtain information about the most distinguishing dimensions between BGs without modelling or additional hypothesis. The underlying structures in the data that explain the most extreme variation between BGs can therefore be revealed in an easily-implemented and interpretable way, and our results show that the MFA components could not be obtained in an intuitive way just "looking" at the correlation matrix.

Another advantage of these techniques is the possibility, at every step on the way in constituting the classification (that is, post-MFA or post-HC) of returning to the variables in order to gain a clearer interpretation of the results (knowing the correlation or contribution of each variable to the components, or having the variables' distributions in each category) and to identify which of them could be a leverage or priority target for public action.

Use of the classification, especially when displayed on a map, makes it possible to quickly discern the exposure profile of a BG and then to obtain more details about its particularities. This type of process could be especially useful for screening purposes, to locate 'hotspots' of strong environmental pressure. Standing alone, it can help policymakers or stakeholders gain insight into cumulative exposure in a given area and adapt their urban planning accordingly. If the data is periodically made available and updated, this approach could also help assess the temporal evolution of exposure profiles, either looking at which category the BG belongs to, or at changes in the profiles themselves.

In addition, the different exposure profiles created following this approach can provide useful information where researchers plan to select a population sample for research on various environmental health issues. Epidemiologists or sociologists, for instance, might wish to investigate in areas having different patterns of environmental exposure, and could then use this approach to easily locate these areas.

There are certain noteworthy precautions to be used with this approach. Firstly, both MFA and HC organize BGs in relation to one another rather than in an absolute way. Therefore, if the same technique is used separately on two different areas, the results are not comparable between these areas, but only within them. This also implies that to use this procedure in a new location or time, it must be entirely repeated on this new particular setting. However, we believe that this approach allows local actors to gain insight into the actual profile of a specific area.

A second limitation (specific to our example though not to the technique) is the relatively small number of environmental exposures included in our analysis, mainly due to the practical, including financial reasons that limited the number of data that could be obtained at the BG scale in the study area. Other air pollutants (particulate matter or ozone, for example), surface or one-off sources (waste disposal or water pollutant industries, for example) could add information and complete assessment of a neighbourhood environmental burden. Obtaining high-quality environmental data at a small geographic scale is an important issue in which there is still room for improvement. Also relative to our examples, the study of the links between the cumulative exposure index and SES or infant mortality require more complex models which take into account confounders and the spatial structure of our data (43).

Finally, this approach could be extended to the wider purpose of assessing populations' living environment by including several other variables, such as accessibility of public transport, primary goods stores, health centres, health professionals, and so on.

Conclusions

Statistical data analysis techniques such as Multiple Factor Analysis and Hierarchical Clustering allow data-driven exploration and classification of the census Block Groups across

different cumulative exposure profiles. Although it cannot be easily used for a formal risk assessment, this approach is able to simultaneously take into account different types of variables and gain insight into the various exposure profiles in an area, in order to reveal where a higher environmental burden exists. Researchers can then investigate more precisely areas with different cumulative exposure profiles.

In order to assess more thoroughly the performance and the value of this approach, particularly as a screening tool for stakeholders, we recommend to test it on areas of different sizes (e.g. municipality, region, state ...) and natures (e.g. cities with other environmental characteristics than the ones explored in the present study or in rural areas). It might also be extended to other ”at risk” exposures (e.g. ambient air particulate matter, pesticides, heavy metals) or to “healthy” exposures, although this could be difficult according. One limitation in this extended assessment is the availability of environmental databases at an appropriate spatial scale. .

Supplementary information is available at *Journal of Exposure Science and Environmental Epidemiology*’s website.

1. WHO. Air Quality Guidelines Global Update 2005 [Internet]. Bonn, Germany: WHO Regional Office for Europe; 2005. Disponible sur: http://www.who.int/phe/health_topics/outdoorair/outdoorair_aqg/en/index.html

2. Barceló MA, Saez M, Saurina C. Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain. Sci Total Environ. 15 oct 2009;407(21):5501‑5523.

3. Brender JD, Maantay JA, Chakraborty J. Residential Proximity to Environmental Hazards and Adverse Health Outcomes. American Journal of Public Health. déc 2011;101(S1):S37‑S52.

4. Passchier-Vermeer W, Passchier WF. Noise exposure and public health. Environ Health Perspect. mars 2000;108(Suppl 1):123‑131.

5. WHO. Guidelines for community noise. 1999 [Internet]. World Health Organization,: Geneva; 2008. Disponible sur: http://www.who.int/docstore/peh/noise/guidelines2.html

6. Mitchell R, Popham F. Greenspace, urbanity and health: relationships in England. J Epidemiol Community Health. 8 janv 2007;61(8):681‑683.

7. Maas J, Verheij RA, Vries S de, Spreeuwenberg P, Schellevis FG, Groenewegen PP. Morbidity is related to a green living environment. J Epidemiol Community Health. 12 janv 2009;63(12):967‑973.

8. Callahan MA, Sexton K. If Cumulative Risk Assessment Is the Answer, What Is the Question? Environ Health Perspect. mai 2007;115(5):799‑806.

9. Pearce JR, Richardson EA, Mitchell RJ, Shortt NK. Environmental justice and health: A study of multiple environmental deprivation and geographical inequalities in health in New Zealand. Social Science & Medicine. août 2011;73:410‑420.

10. Sexton K. Cumulative Risk Assessment: An Overview of Methodological Approaches for Evaluating Combined Health Effects from Exposure to Multiple Environmental Stressors. International Journal of Environmental Research and Public Health. 26 janv 2012;9(12):370‑390.

11. Alves S, Tilghman J, Rosenbaum A, Payne-Sturges DC. U.S. EPA authority to use cumulative risk assessments in environmental decision-making. Int J Environ Res Public Health. juin 2012;9(6):1997‑2019.

12. U.S. EPA. Framework for Cumulative Risk Assessment [Internet]. Washington DC: U.S. Environmental Protection Agency; 2003 mai. Report No.: EPA/630/P02/001F. Disponible sur: http://www.epa.gov/raf/publications/framework-cra.htm

13. California Environmental Justice Action Plan. Environmental Justice Action Plan [Internet]. Cal-EPA; 2004 oct. Disponible sur: http://www.calepa.ca.gov/EnvJustice/ActionPlan/Documents/October2004/ActionPlan.pdf

14. Commission of the European Communities. The European Environment & Health Action Plan 2004-2010 COM(2004) 416 final [Internet]. Commission of the European Communities; 2004. Disponible sur: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52004DC0416:EN:HTML

15. National Research Council (U.S.). Science and decisions advancing risk assessment. Washington, D.C.: National Academies Press; 2009.

16. Su JG, Morello-Frosch R, Jesdale BM, Kyle AD, Shamasunder B, Jerrett M. An index for assessing demographic inequalities in cumulative environmental hazards with application to Los Angeles, California. Environ Sci Technol. 15 oct 2009;43(20):7626‑7634.

17. Richardson EA, Mitchell R, Shortt NK, Pearce J, Dawson TP. Developing summary measures of health-related multiple physical environmental deprivation for epidemiological research [Abstract only]. Environment and Planning A. 2010;42(7):1650‑1668.

18. Yorita Christensen KL, White P. A methodological approach to assessing the health impact of environmental chemical mixtures: PCBs and hypertension in the National Health and Nutrition Examination Survey. Int J Environ Res Public Health. nov 2011;8(11):4220‑4237.

19. Huang G, London JK. Cumulative Environmental Vulnerability and Environmental Justice in California's San Joaquin Valley. International Journal of Environmental Research and Public Health. 3 mai 2012;9(5):1593‑1608.

20. INSEE. Institut national de la statistique et des études économiques [Internet]. Disponible sur: http://www.insee.fr

21. Soulhac L, Salizzoni P, Cierco F-X, Perkins R. The model SIRANE for atmospheric urban pollutant dispersion; part I, presentation of the model. Atmospheric Environment. déc 2011;45(39):7379‑7395.

22. Soulhac L, Salizzoni P, Mejean P, Didier D, Rios I. The model SIRANE for atmospheric urban pollutant dispersion; PART II, validation of the model on a real case study. Atmospheric Environment. mars 2012;49:320‑337.

23. Padilla CM, Deguen S, Lalloue B, Blanchard O, Beaugard C, Troude F, et al. Cluster analysis of social and environment inequalities of infant mortality. A spatial study in small areas revealed by local disease mapping in France. Science of The Total Environment. 1 juin 2013;454–455:433‑441.

24. European Parliament and European Council. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Communities. 2008;

25. European Parliament and European Council. Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise. Official Journal of the European Communities. 2002;

26. Kihal-Talantikite W, Padilla CM, Lalloue B, Rougier C, Defrance J, Zmirou-Navier D, et al. An exploratory spatial analysis to assess the relationship between deprivation, noise and infant mortality: an ecological study. Environmental Health. 16 déc 2013;12(1):109.

27. E-PRTR - The European Pollutant Release and Transfer Register [Internet]. Disponible sur: http://prtr.ec.europa.eu/

28. Basias - Inventaire historique de sites industriels et activités de service [Internet]. Disponible sur: http://basias.brgm.fr/

29. ESRI ADDG. ArcGIS 9.1. The complete geographic information system. 2005;

30. Kihal-Talantikite W, Padilla CM, Lalloué B, Gelormini M, Zmirou-Navier D, Deguen S. Green space, social inequalities and neonatal mortality in France. BMC Pregnancy and Childbirth. 20 oct 2013;13(1):191.

31. Lebart L, Morineau A, Warwick KM. Multivariate descriptive statistical analysis : correspondence analysis and related techniques for large matrices. New York: Wiley; 1984.

32. Lalloué B, Monnez J-M, Padilla C, Kihal W, Le Meur N, Zmirou-Navier D, et al. A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis. International Journal for Equity in Health. 2013;12(1):21.

33. Escofier B, Pagès J. Multiple factor analysis (AFMULT package). Computational Statistics & Data Analysis. août 1994;18(1):121‑140.

34. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Disponible sur: http://www.R-project.org/

35. Lê S, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. Journal of statistical software. 2008;25(1):1‑18.

36. O'Neill MS, Jerrett M, Kawachi I, Levy JI, Cohen AJ, Gouveia N, et al. Health, wealth, and air pollution: advancing theory and methods. Environ Health Perspect. déc 2003;111(16):1861‑1870.

37. Sexton K, Hattis D. Assessing cumulative health risks from exposure to environmental mixtures - three fundamental questions. Environ Health Perspect. mai 2007;115(5):825‑832.

38. Menzie CA, MacDonell MM, Mumtaz M. A phased approach for assessing combined effects from multiple stressors. Environ Health Perspect. mai 2007;115(5):807‑816.

39. Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the Health Effects of Exposure to Multi-Pollutant Mixture. Annals of Epidemiology. févr 2012;22(2):126‑141.

40. Benmarhnia T, Laurian L, Deguen S. Measuring Spatial Environmental Deprivation: A New Index and its Application in France. Environmental Justice. avr 2013;6(2):48‑55.

41. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. International journal of epidemiology. 1989;18(1):269‑74.

42. Shortt NK, Richardson EA, Pearce J, Mitchell RJ. Mortality inequalities by environment type in New Zealand. Health & Place. sept 2012;18(5):1132‑1136.

43. Deguen S, Lalloué B, Bard D, Havard S, Arveiler D, Zmirou-Navier D. A small-area ecologic study of myocardial infarction, neighborhood deprivation, and sex: a Bayesian modeling approach. Epidemiology. juill 2010;21(4):459‑466.

FIGURES LEGEND

Figure 1.Spatialdistribution of several environmental exposures and of the cumulative exposure categories for the Grand Lyon area (France).

Figure 2.Overview of the methodological procedure used to create the cumulative exposure categories.

Figure 3.Distribution of the contextual socio-economic index by cumulative exposure categories.

SUPPLEMENTARY FIGURES LEGEND

Figure S1. Circle of correlations of the Multiple Factor Analysis (quantitative variables).

Figure S2. Projection of the qualitative variables' modalities on the 2 first axes of the Multiple Factor Analysis.

Figure S3. Spatial distribution of the fourth first factors of the MFA (categories in quintiles).

Table 1. Pearson Correlations Between the Selected Quantitative Environmental Exposures

	NO2_2003	NO2_2006	Noise_arithmetic	Noise_median	Noise_pop_energetic	Noise_pop_arithmetic	Traffic_200m	Traffic_250m	Soil_count	Soil_500m	Soil_1000m	Greenspaces_area	Greenspaces_index
NO2_2003	1												
NO2_2006	1.00	1											
Noise_arithmetic	0.23	0.24	1										
Noise_median	0.32	0.32	0.92	1									
Noise_pop_energetic	0.18	0.18	0.20	0.21	1								
Noise_pop_arithmetic	0.17	0.17	0.26	0.27	0.99	1							
Traffic_200m	0.58	0.58	0.17	0.24	0.27	0.26	1						
Traffic_250m	0.53	0.53	0.17	0.22	0.29	0.29	0.97	1					
Soil_count	0.23	0.23	0.18	0.21	0.11	0.11	0.17	0.17	1				
Soil_500m	0.49	0.49	0.13	0.28	0.14	0.13	0.37	0.34	0.64	1			
Soil_1000m	0.57	0.57	0.11	0.28	0.12	0.12	0.40	0.35	0.52	0.94	1		
Greenspaces_area	-0.41	-0.41	-0.08	-0.12	-0.16	-0.18	-0.45	-0.47	-0.09	-0.21	-0.25	1	
Greenspaces_index	-0.28	-0.27	0.02	-0.05	-0.08	-0.10	-0.38	-0.38	-0.20	-0.28	-0.28	0.61	1

Table 2. Descriptive Statistics of the Selected Quantitative Environmental Variables in the LMA

Type	Variables	Min	Median	Mean	Max
Air pollution (NO_2)	2003 average ($\mu g/m^3$)	29.1	40.2	40.3	59.7
	2006 average ($\mu g/m^3$)	32.7	42.2	42.4	60.7
Noise (L_{DEN})	Arithmetic mean (dB(A))	53.0	63.9	63.9	74.3
	Median (dB(A))	51.3	64.2	64.2	76.8
	Energetic mean with population (dB(A))	0.0	68.6	67.1	79.6
	Arithmetic mean with population (dB(A))	0.0	65.9	64.2	76.9
Polluted sites	Number	0	0	0.9	12
	Number of 500m buffers	0	3	7.1	38
	Number of 1000m buffers	0	9	17.8	74
Traffic proximity	Population within 200m (%)	0.0	89.9	77.5	100
	Population within 250m (%)	0.0	98.9	83.4	100
Green spaces	Green spaces area (km^2)	0	0.02	0.19	3.89
	Green spaces proportion (%)	0	7.4	11.9	57.0

Table 3. Descriptive Statistics of the Selected Qualitative Environmental Variables in the LMA

Type	Variables	**0**[1]	**1**[1]	**2 or 2+**[1]
Industrial proximity	Number	480 (96.2%)	13 (2.6%)	6 (1.2%)
	Presence of 500m buffer	422 (84.6%)	77 (15.4%)	-
	Number of 500m buffers	422 (84.6%)	56(11.2%)	21 (4.2%)
	Presence of 1km buffer	356 (71.3%)	143 (28.7%)	-

[1]number of BGs with the modality (% of BG with the modality)

Table 4. Contributions of Environmental Groups to the Four First Components of the MFA

Groups	Component 1	Component 2	Component 3	Component 4
Air Pollution	27.4	3.1	0.7	2.1
Noise	10.3	14.5	45.6	52.8
Industrial Proximity	0.1	76.8	2.7	8.5
Traffic Proximity	26.1	0.7	3.9	4.7
Polluted sites	18.8	3.7	6.6	29.4
Green Spaces	17.4	1.2	40.6	2.5

Contribution of the group to the component (in %)

Table 5. Average Values of the Quantitative Variables per Category of Cumulative Exposure Created with HC

Variables	Category 1 *(n=13)*	Category 2 *(n=36)*	Category 3 *(n=186)*	Category 4 *(n=70)*	Category 5 *(n=194)*	Total *(n=499)*
NO2_2003 (µg/m^3)	39.9	32.5	36.9	39.6	45.3	40.3
NO2_2006 (µg/m^3)	42.0	35.1	39.2	41.8	47.2	42.4
Noise_arithmetic (dB(A))	63.0	62.7	63.2	65.7	64.2	63.9
Noise_median (dB(A))	63.0	62.5	63.0	65.8	65.2	64.2
Noise_pop_energetic (dB(A))	0.0	67.8	66.9	69.5	70.9	67.1
Noise_pop_arithmetic (dB(A))	0.0	63.9	64.1	67.4	67.5	64.2
Traffic_200m (%)	52.8	35.6	65.9	83.0	96.1	77.5
Traffic_250m (%)	53.8	43.4	75.3	89.3	98.6	83.4
Soil_count	0.5	0.5	0.3	1.0	1.7	1.0
Soil_500m	5.2	1.7	1.8	5.5	13.9	7.1
Soil_1000m	16.4	3.2	5.3	12.1	34.7	17.8
Greenspaces_area (km²)	0.61	1.42	0.11	0.13	0.03	0.19
Greenspaces_index (%)	16.5	34.1	13.0	8.5	7.6	11.9

n: number of census blocks groups in the category

Table 6. Distribution of the Modalities of the Qualitative Variables per Cumulative Exposure Category Created with HC

Variables	Modalities	Category 1 *(n=13)*	Category 2 *(n=36)*	Category 3 *(n=186)*	Category 4 *(n=70)*	Category 5 *(n=194)*	Total *(n=499)*
Industries_count	0	12 *(2.5%)* *92.3%*	35 *(7.3%)* *97.2%*	186 *(38.8%)* *100%*	53 *(11.0%)* *75.7%*	194 *(40.4%)* *100%*	480 *96.2%*
	1	0 *(0%)* *0%*	1 *(7.7%)* *2.8%*	0 *(0%)* *0%*	12 *(92.3%)* *17.1%*	0 *(0%)* *0%*	13 *2.6%*
	2	1 *(16.7%)* *7.7%*	0 *(0%)* *0%*	0 *(0%)* *0%*	5 *(83.3%)* *7.1%*	0 *(0%)* *0%*	6 *1.2%*
Industries_500m_presence	0	10 *(2.4%)* *76.9%*	33 *(7.8%)* *91.7%*	185 *(43.8%)* *99.5%*	0 *(0%)* *0%*	194 *(46.0%)* *100%*	422 *84.6%*
	1	3 *(3.9%)* *23.1%*	3 *(3.9%)* *8.3%*	1 *(1.3%)* *0.5%*	70 *(91.0%)* *100%*	0 *(0%)* *0%*	77 *15.4%*
Industries_500m_count	0	10 *(2.4%)* *76.9%*	33 *(7.8%)* *91.7%*	185 *(43.8%)* *99.5%*	0 *(0%)* *0%*	194 *(46.0%)* *100%*	422 *84.6%*
	1	2 *(3.6%)* *15.4%*	2 *(3.6%)* *5.6%*	0 *(0%)* *0%*	52 *(92.8%)* *74.3%*	0 *(0%)* *0%*	56 *11.2%*
	2+	1 *(4.8%)* *7.7%*	1 *(4.8%)* *2.8%*	1 *(4.8%)* *0.5%*	18 *(85.7%)* *25.7%*	0 *(0%)* *0%*	21 *4.2%*
Industries_1km_presence	0	10 *(2.8%)* *76.9%*	29 *(8.2%)* *80.6%*	146 *(41.0%)* *78.5%*	0 *(0%)* *0%*	171 *(48.0%)* *88.1%*	356 *71.3%*
	1	3 *(2.1%)* *23.1%*	7 *(4.9%)* *19.4%*	40 *(28.0%)* *21.5%*	70 *(49.0%)* *100%*	*23 (16.1%)* *11.9%*	143 *28.7%*

n *(x%)* : n is the number of BGs with the modality. x% is the percentage of BGs in the category among all the BGs with the modality.
y% y% is the percentage of BGs with the modality among the BGs of the category

Table 7. Average infant mortality rates (in ‰) at the BG level, according to several environmental categorization

		Average infant mortality rate (‰)	*p-value*[4]
Cumulative exposure categories	Category 1	2.10	$<10^{-15}$
	Category 2	2.97	
	Category 3	4.78	
	Category 4	4.18	
	Category 5	3.96	
Quintiles of NO_2 concentration levels[1]	NO2_1	3.59	0.18
	NO2_2	4.52	
	NO2_3	4.31	
	NO2_4	5.15	
	NO2_5	3.38	
Quintiles of noise levels[2]	Noise_1	4.34	0.65
	Noise_2	4.52	
	Noise_3	4.41	
	Noise_4	3.98	
	Noise_5	3.75	
Quintiles of green spaces areas[3]	Greenspaces_1	4.54	$<10^{-2}$
	Greenspaces_2	4.93	
	Greenspaces_3	4.37	
	Greenspaces_4	3.83	
	Greenspaces_5	3.35	

NB : due to the different compositions of the categories, comparisons between exposures are impossible

[1] Using NO2_2006

[2] Using Noise_pop_energetic

[3] Using Greenspaces_index

[4] p-value for the Kruskal-Wallis test

Table S1. Variance Explained by the Five First Components of the MFA on the Selected Variables

	% of variance explained	Cumulative % of variance
Component 1	29.7	29.7
Component 2	13.7	43.4
Component 3	10.7	54.1
Component 4	9.4	63.5
Component 5	7.2	70.7

Table S2. Correlations and Contributions of Variables to the Four First Components of the MFA

Variables		Component 1		Component 2		Component 3		Component 4	
		Coord [1]	*Ctr* [2]	*Coord* [1]	*Ctr* [2]	*Coord* [1]	*Ctr* [2]	*Coord* [1]	*Ctr* [2]
NO2_2003		0.82	13.76	-0.19	1.54	0.07	0.30	0.13	1.06
NO2_2006		0.82	13.65	-0.19	1.58	0.08	0.38	0.12	1.00
Noise_arithmetic		0.34	1.93	0.43	6.92	0.59	16.02	0.05	0.12
Noise_median		0.44	3.32	0.35	4.54	0.58	15.96	0.10	0.53
Noise_pop_energetic		0.38	2.46	0.18	1.14	0.37	6.53	-0.71	26.52
Noise_pop_arithmetic		0.39	2.54	0.23	1.87	0.39	7.03	-0.70	25.61
Traffic_200m		0.80	13.39	0.07	0.20	-0.18	1.80	-0.16	1.77
Traffic_250m		0.78	12.67	0.10	0.45	-0.19	2.08	-0.21	2.96
Soil_count		0.42	2.98	-0.03	0.03	0.25	2.86	0.42	9.44
Soil_500m		0.67	7.56	-0.19	1.33	0.21	2.16	0.45	10.73
Soil_1000m		0.70	8.24	-0.25	2.38	0.18	1.57	0.42	9.21
Greenspaces_area		-0.63	9.92	-0.05	0.16	0.49	17.17	0.18	2.51
Greenspaces_index		-0.54	7.43	-0.14	1.04	0.58	23.39	-0.02	0.03
Industries_count	0	0.02	0.00	-0.11	0.35	0.00	0.00	-0.04	0.08
	1	-0.70	0.07	2.72	5.32	-0.10	0.01	0.61	0.56
	2	0.23	0.00	3.27	3.55	-0.01	0.00	1.66	1.92
Industries_500m_presence	0	0.02	0.00	-0.40	3.76	0.05	0.10	-0.09	0.37
	1	-0.11	0.01	2.20	20.61	-0.28	0.54	0.48	2.03
Industries_500m_count	0	0.02	0.00	-0.40	3.76	0.05	0.10	-0.09	0.37
	1	-0.07	0.00	2.07	13.25	-0.45	1.05	0.34	0.74
	2+	-0.21	0.01	2.55	7.55	0.19	0.07	0.85	1.77
Industries_1km_presence	0	0.04	0.01	-0.52	5.35	0.09	0.25	-0.07	0.19
	1	-0.10	0.02	1.30	13.33	-0.22	0.62	0.17	0.48

[1] Coordinate of the variable on the axis, i.e correlation coefficient with the component for quantitative variables

[2] Contribution of the variable or modality to the component (in %)

Text S1: Principal component analysis, Multiple factor analysis and Hierarchical clustering

Principal component analysis[1–3]

Principal component analysis (PCA) is a data analysis technique which aims to describe (highlight the similarity and dissimilarity between the statistical units and the correlations between the variables), summarize (determine a small number of new variables, uncorrelated linear combinations of the originals with maximal variance) and visualize the information contained in a data set.

Let *X* be a dataset with *n* rows representing the statistical units and *p* columns representing the variables (with the mean of each variable to zero). x_i^j will then be the value of the variable *j* for the unit *i*. Let s^j be the standard deviation of the j^{th} variable. Let *D* be a diagonal matrix for the weights of the statistical units (frequently all the weights are equal to *1/n*).

PCA first studies the statistical units in the space of variables' with the purpose to find a viewable graphical representation of these units, such as units with similar values will be represented by close points and units with very different values will be represented by distant points. To do so, a mathematical distance must be chosen. A very commonly used distance between the points *i* and *i'* is $d(i,i') = \sqrt{\sum_{j=1}^{p} \frac{1}{(s^j)^2}(x_i^j - x_{i'}^j)^2}$ (in the case of normed PCA), but other distances can be used according to the purpose. The distance *d* is equivalent to set as metric on the space of the statistical units the diagonal matrix *M* such as $M_{j,j} = \frac{1}{(s^j)^2}$.

Then, the aim of CA is to visualize these points. Since the space's dimension (i.e. the number of variables) is in general high, it is necessary to project the points on an optimal subspace of lower dimension, in order to have the most precise and faithful representation of the initial

scatter plot. Hence, PCA determines a subspace F_r (with dimension r) which maximizes the inertia of the projected points on F_r with respect to the barycenter G (i.e. the weighted sum of the squared distances between G and the projections on F_r). To do so, an iterative process is used and leads to determine a set of orthonormal vectors $(\overrightarrow{u_1}, \ldots, \overrightarrow{u_k}, \ldots, \overrightarrow{u_r})$ which constitute a base of F_r, where $\overrightarrow{u_k}$ is the eigenvector of the matrix X^TDXM corresponding to the k^{th} largest eigenvalue, λ_k. The axis $(G, \overrightarrow{u_k})$ is called the k^{th} *principal axis* and you can get up to p principal axes, i.e the number of original variables. At this step, it is already possible to obtain the quality of representation and the contribution of each point on each principal axis.

Statistically, a principal axis represents a linear combination of the original variables called *principal component* and can be interpreted as a linear combination with maximal variance (equal to λ_k for the k$^{\text{th}}$ principal component) given the constraint to be uncorrelated with the previous components. The ratio $\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p}$ can be statistically interpreted as the percentage of the total variance explained by the k^{th} component. By construction, components are ordered from the one which explains the higher proportion of total variance to the one which explains the lower.

The interpretation of the PCA follows different steps. First, the number of axes to keep must be chosen. Secondly, thanks to the correlation circle created by the PCA, it is possible to analyze the correlations between the variables and to interpret the principal components by studying their correlations with the variables. The last step is the interpretation of the position of the points representing the units using their projections on the principal planes.

Multiple factor analysis[4]

When the set of variables consists of several groups, each with a different number of variables, PCA can be inappropriate since a group with a lot of variables can dominate groups with fewer variables. Then, a solution is to give in some sense an equal weight to each group (rather than to each variable as in the case of normed PCA) regardless to the number of variables in it. This is the purpose of the multiple factor analysis (MFA). Another advantage of MFA is its ability to take into account at the same time groups of quantitative variables and groups of qualitative variables.

Let $X = (X^1 | \dots | X^q)$ be a data set composed of observations among the same units of q groups of variables, and denote m_k the number of variables in the subset X^k.

The first step of the MFA is, for each subset $X^k, k = 1,2, \dots q$, to perform a PCA if the group is composed of quantitative variable or a multiple correspondence analysis (MCA) if the group is composed of qualitative (dummy) variables (MCA is a very similar technique to PCA but for qualitative variables). For $k = 1,2, \dots q$, let M^k be the metric for analysis of group *k*, i.e. the matrix used to compute distances between units (in the case of a group of quantitative variables, where data is mean-centred and standardized, this metric is the identity matrix for a normed PCA), and let λ_1^k be the highest eigenvalue of the analysis on group *k*.

The second step of the MFA is to perform a PCA on the dataset *X* with the metric *M*, where *M* is the block-diagonal matrix $M = \begin{pmatrix} \frac{M^1}{\lambda_1^1} & & \\ & \ddots & \\ & & \frac{M^q}{\lambda_1^q} \end{pmatrix}$. This allows in some sense the same weight to be given to each group, even if they are of very different sizes, and comparison of groups of different types of variables.

The interpretation of the MFA is then similar to a PCA: new variables are created as linear combinations of the original variables (without any factors rotation) and can be interpreted using correlations or contributions together with a set of graphical outputs which facilitate interpretation.

Hierarchical clustering[3,5]

Hierarchical clustering (HC) is an unsupervised method of clustering which creates a hierarchy of classes (i.e. clusters), frequently used after a PCA or other data analysis techniques such as MFA.

Let I be a set of n elements ($I = \{1,2, \dots, n\}$) represented by points in $\mathbb{R}^p$ and d a distance between points. The purpose is to find a partition in r classes which maximizes the between-classes inertia or, which is equivalent, which minimizes the within-classes inertia. This clustering criterion based on inertia allows creating classes homogeneous in their composition and heterogeneous between them. In practice, the search for a direct optimal solution requires generally too many computations and an approximation must be used. To do so, we use the algorithm of hierarchical clustering with a particular distance Δ between classes (based on d). This distance, called Ward's distance, is defined as follows: let I_1 and I_2 be two classes, p_1 and p_2 their respective weights, and G_1 and G_2 their respective barycenters, then the Ward's distance between classes I_1 and I_2 is $\Delta(I_1, I_2) = \frac{p_1 p_2}{p_1 + p_2} d^2(G_1, G_2)$.

The algorithm of hierarchical (ascending) clustering is then:

- Step1: from the partition containing all the singletons $P_0 = \{\{1\}, \{2\}, \dots, \{n\}\}$ the distance Δ is computed for all the pairs of singletons. Classes $\{l\}$ and $\{m\}$ with minimum distance Δ are merged and then the partition with *(n-1)* elements obtained is

$P_1 = \{\{l, m\}, \{1\}, \ldots, \{n\}\}$. In other words, the two closest elements of *I* are merged in a single class while all the others remain singletons.

- …
- Step *r*: from the partition P_{r-1} with *(n-(r-1))* elements, the distance Δ is computed between all the elements of the partition. Minimal distance classes are merged and the partition with *(n-r)* elements, P_r, is then created. The merging of these two classes is, by definition of the Ward's distance, the one which minimizes the loss of between-classes inertia among all the other possible mergings at this step.
- …
- Step *n-1*: the partition created is $P_{n-1} = \{I\}$

The HC presents as results a dendrogram (illustrating for each step of the algorithm the loss of between-classes inertia). The first step in the interpretation of the results is to choose the number of classes to keep. Generally, the partition that is chosen is the one preceding a strong decrease in the between classes inertia. However, other partitions can be chosen according to the purpose of the clustering. Once the number of classes determined, it is possible to interpret each class thanks to the comparison of the descriptive statistics of the variables between the class and the whole set.

References

1. Hastie T, Tibshirani R, Friedman J. Principal Components. In: *The elements of statistical learning*. 2nd ed. Springer; 2009:534‑541.

2. Jolliffe IT. *Principal component analysis*. 2nd ed. New York: Springer; 2002.

3. Lebart L, Morineau A, Warwick KM. *Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices*. New York: Wiley; 1984.

4. Escofier B, Pagès J. Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis*. 1994;**18**(1):121‑140.

5. Hastie T, Tibshirani R, Friedman J. Hierarchical Clustering. In: *The elements of statistical learning*. 2nd ed. Springer; 2009:520‑528.